UNIVERSITÀ DEGLI STUDI DI PADOVA

quantexa
Connecting Data | Empowering Decisions

A CNN-RNN Framework for Image Annotation from Visual Cues and Social Network Metadata

Tobia Tesan[1], Pasquale Coscia[2] and Lamberto Ballan[2]
[1] Quantexa Ltd, London, UK
[2] University of Padova, Department of Mathematics, Italy
Visual Intelligence and Machine Perception (VIMP) Group

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Image annotation

- Process of labelling images using text or annotation tools.
- Some images might be hard to recognize without additional context.
- Weakly-annotated images may help to disambiguate the visual classification task.

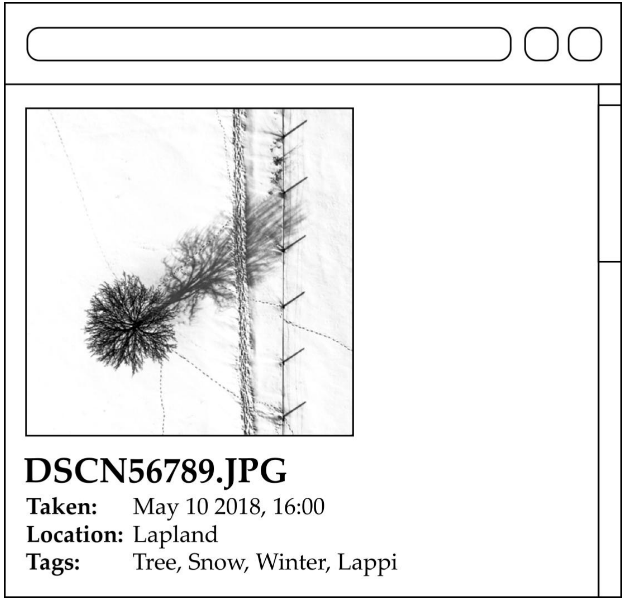
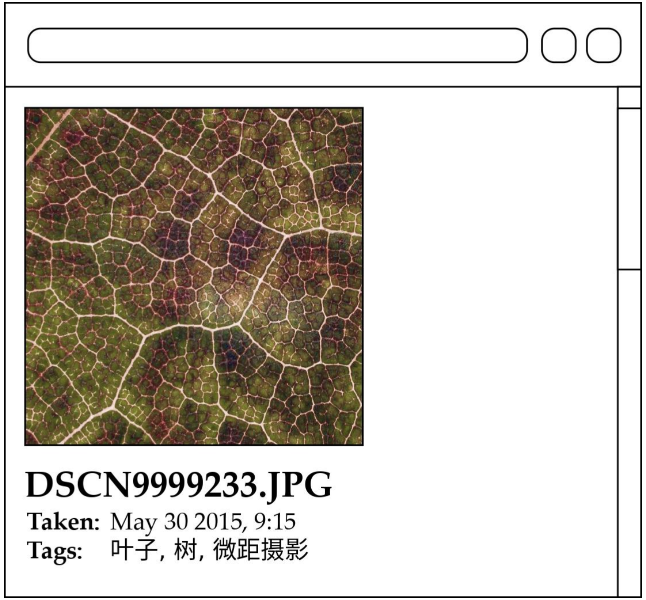
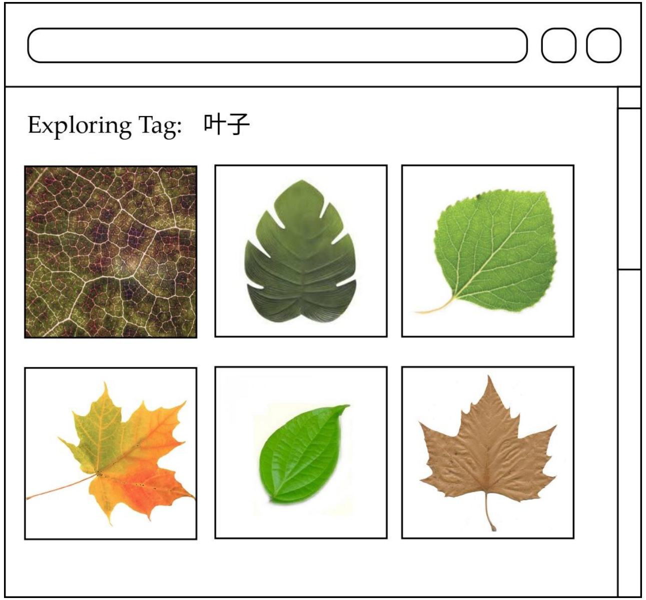


Image annotation

- Metadata of images shared on social-media are an ideal source of additional information.





Metadata Limitations

- Image metadata are useful but can be:
  - noisy
  - highly subjective

- Models should also be robust to vocabulary changes.

[0 1 ... 1 1]

[1 1 ... 1 0]

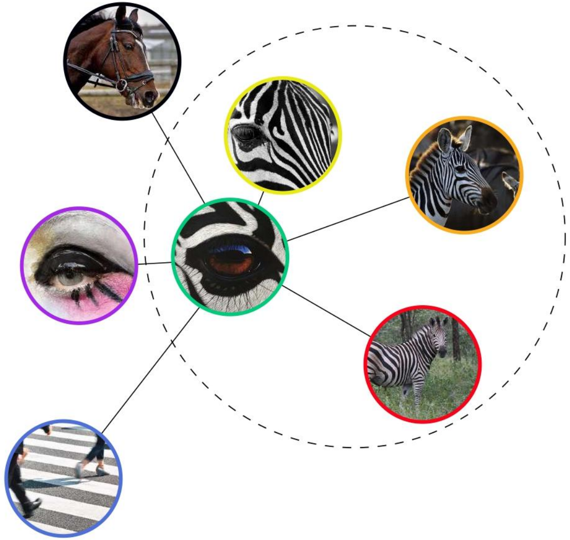
vocabulary = [dog , cat, . . . cute, laptop]

[1 1 ... 1 1]

vocabulary = [dog , cat, . . . cute, laptop, ***beard****]*

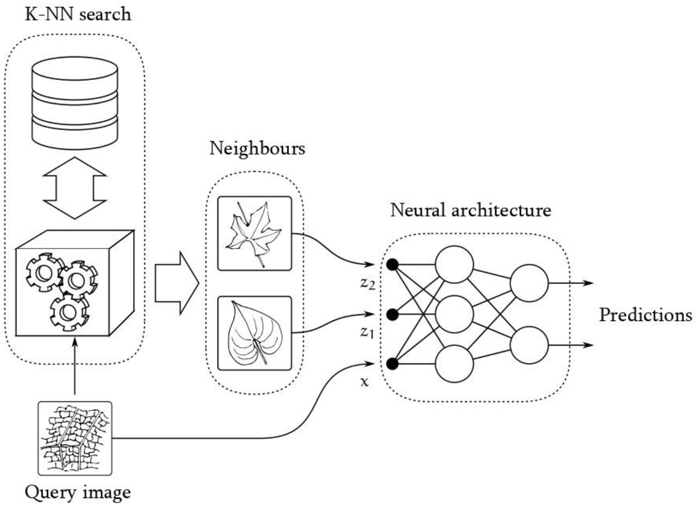
Our approach

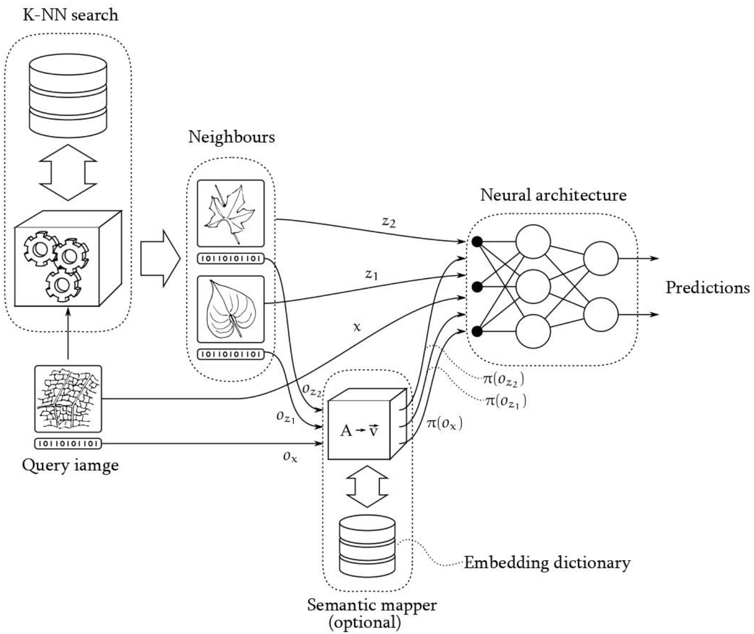
- Advanced semantic mapping and CNN-RNN fusion schemes.
- Visual features and metadata to jointly leverage context and visual cues.
- State-of-the-art results on the multi-label image annotation task using the NUS-WIDE dataset.
- Our models decrease both sensory and semantic gaps to better annotate images.



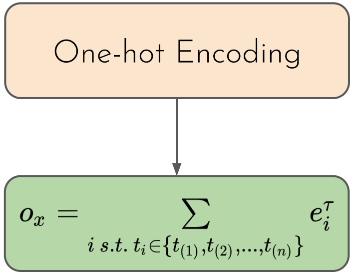
Context (tags) + Visual Cues

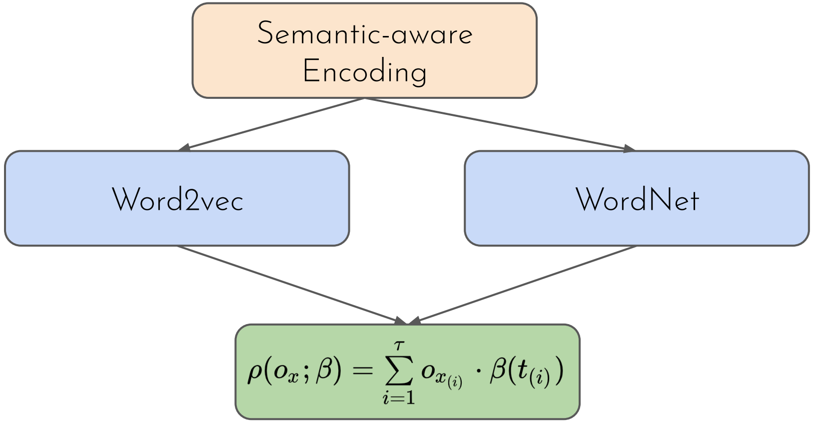
Visual models vs Joint Models



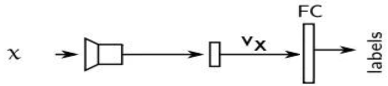


Metadata Encoding

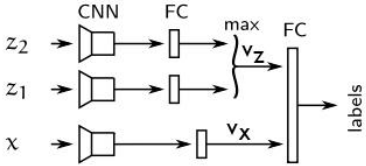




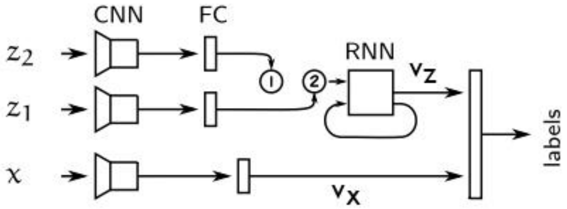
Visual Models



Visual only

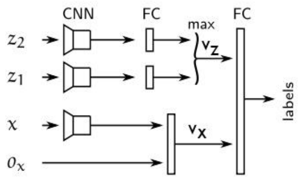


LTN

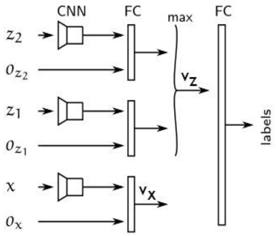


RTN

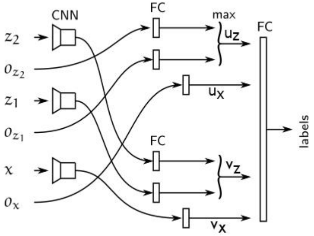
Joint Models



LTN+Vecs

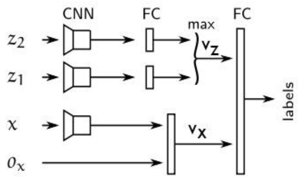


LTN+AllVecs

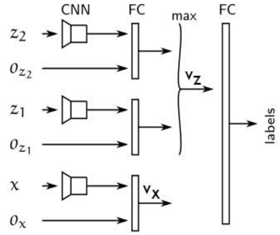


LTwin

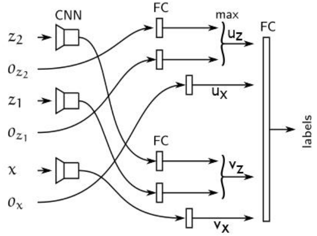
Joint Models



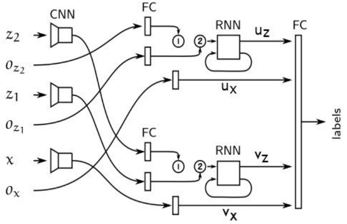
LTN+Vecs



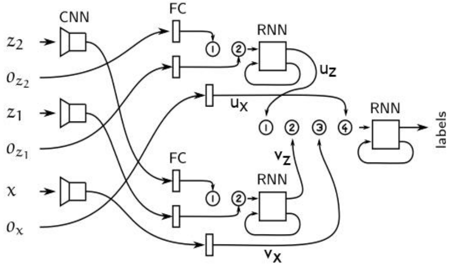
LTN+AllVecs



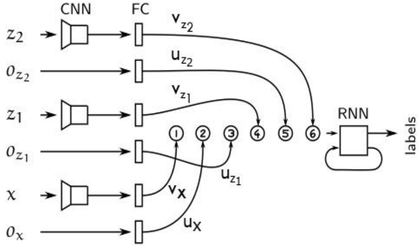
LTwin



LTwin+RNN



LTwin+2RNN



LZip

Dataset & Metrics

- NUS-WIDE dataset:
  - 269,648 images collected from Flickr;
  - 81 labels (manual annotation);
  - 5000 most frequent tags.

- Metrics:
  - Per-label/per-image mean Average Precision (mAP);
  - Precision and recall.

	Image	Label	Metadata (tags)
Image: 163792		grass	centipede, yellow, naturesfinest, k1ood, macro, pentax, kit, eyes, animals, grass, chenille, nature, 1855, johannpix, caterpillar
Neighbour: 140470		animal	flickrdiamond, animalkingdomelite, dragonfly, **naturesfinest***, **k1ood***, **macro***, **pentax***, wild, **kit***, **animals***, blue, damselfly, green, **nature***, blueribbonwinner, **1855***, diamondclassphotographer, closeup, **johannpix***, libellule
Neighbour: 140175		sun, sky, flowers, clouds	**yellow***, **naturesfinest***, **k1ood***, **pentax***, flash, soe, **kit***, outdoors, overtheshot, **1855***, colors, sun, flowers, **johannpix***, sky, tulips, fillin
Neighbour: 15106		animal	**yellow***, **macro***, 5hits, selectivecolorization, **animals***, selectivecolor, bird, **nature***, chicken, chick, beak, baby, bw

NUS-WIDE dataset

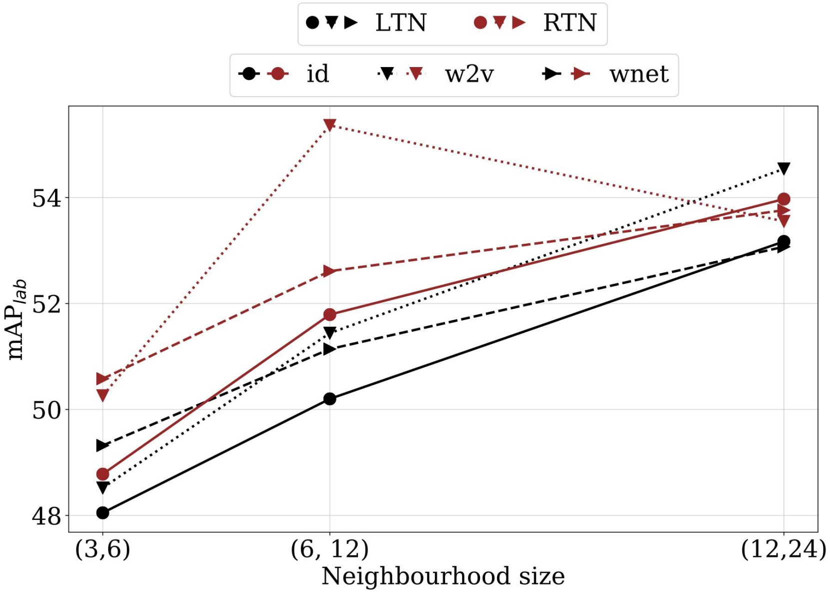
Experimental Results (1/4)

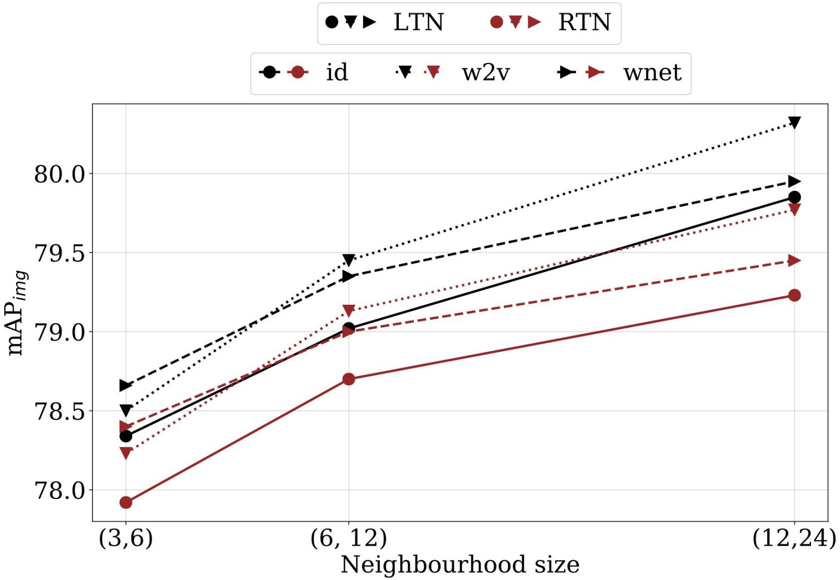
- Our best results in comparison to several baselines and SOTA models.

Method	mAP_{lab}	mAP_{img}	rec_{lab}	$prec_{lab}$	rec_{img}	$prec_{img}$
Tag-only Model + linear SVM [7]	46.67	-	-	-	-	-
Graphical Model (all metadata) [7]	49.00	-	-	-	-	-
CNN + WARP [16]	-	-	35.60	31.65	60.49	48.59
CNN-RNN [21]	-	-	30.40	40.50	61.70	49.90
SR-RNN [22]	-	-	50.17 $\star$	55.65 $\star$	71.35 $\star$	70.57 $\star$
SR-RNN + Vecs [22] †	-	-	58.52 $\star$	63.51 $\star$	77.33 $\star$	76.21 $\star$
SRN [35]	60.00	80.60	41.50 $\star$	70.40 $\star$	58.70 $\star$	81.10 $\star$
MangoNet [33]	62.80	80.80	41.00 $\star$	73.90 $\star$	59.90 $\star$	80.60 $\star$
LTN [2]	52.78 ± 0.34	80.34 ± 0.07	43.61 ± 0.47	46.98 ± 1.01	74.72 ± 0.16	53.69 ± 0.13
LTN + Vecs [2] †	61.88 ± 0.36	80.27 ± 0.08	57.30 ± 0.44	54.74 ± 0.63	75.10 ± 0.20	53.46 ± 0.09
Upper bound	100.00 ± 0.00	100.00 ± 0.00	65.82 ± 0.35	60.68 ± 1.32	92.09 ± 0.10	66.83 ± 0.12
Our baseline: v-only	45.05 ± 0.11	76.88 ± 0.11	42.31 ± 0.59	43.74 ± 1.07	71.41 ± 0.13	51.36 ± 0.13
Our baseline: $LTN_{\mathtt{n:id}}$	53.17 ± 0.12	79.82 ± 0.16	45.67 ± 1.75	47.64 ± 2.18	74.29 ± 0.13	53.34 ± 0.17
Our baseline: LTN + $Vecs_{\mathtt{n:id,f:id}}$ †	54.86 ± 0.20	81.34 ± 0.15	46.56 ± 1.39	50.10 ± 1.70	75.67 ± 0.17	54.37 ± 0.14
Our model: $RTN_{\mathtt{n:w2v}}$	55.36 ± 0.34	79.77 ± 0.27	48.73 ± 2.77	51.21 ± 2.61	74.35 ± 0.29	53.28 ± 0.24
Our model: $LTwin_{\mathtt{n:w2v,f:w2v}}$ †	**63.13** ± 0.31	**83.77** ± 0.06	54.40 ± 1.33	51.86 ± 1.58	78.06 ± 0.05	55.78 ± 0.13

Experimental Results (2/4)

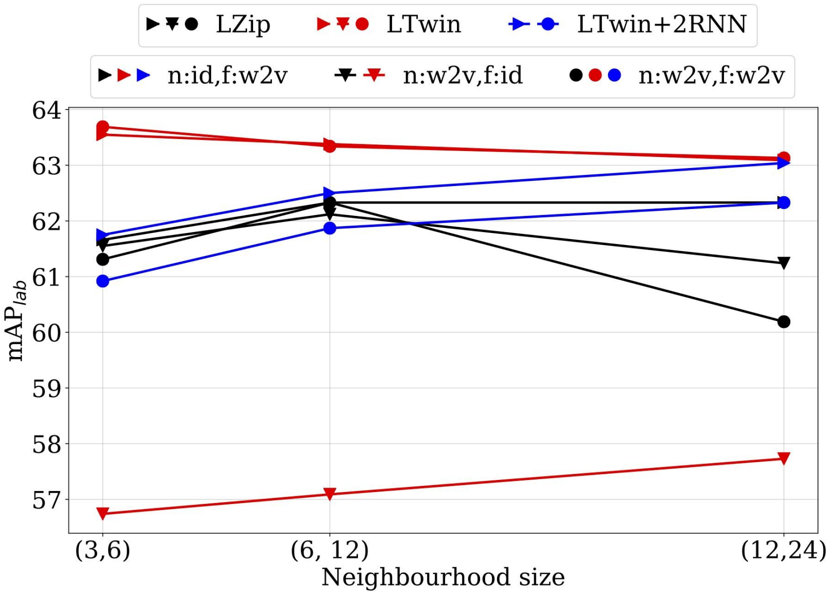
- mAP_{lab} and mAP_{img} for visual models.

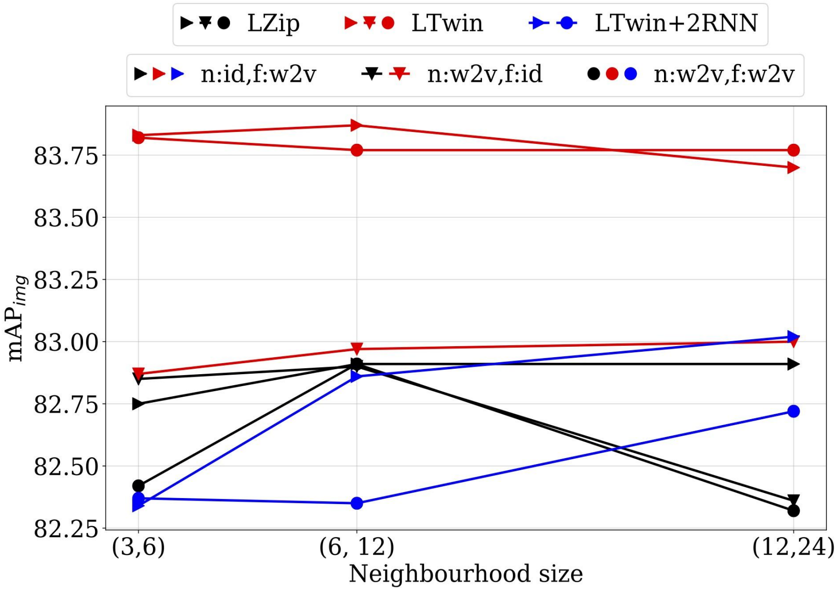




Experimental Results (3/4)

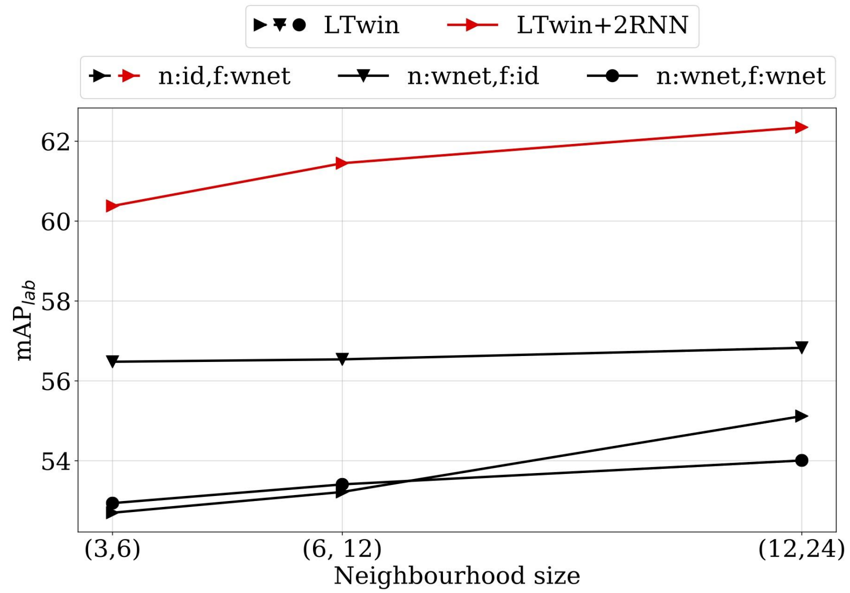
- mAP_{lab} and mAP_{img} for joint models (word2vec embeddings).

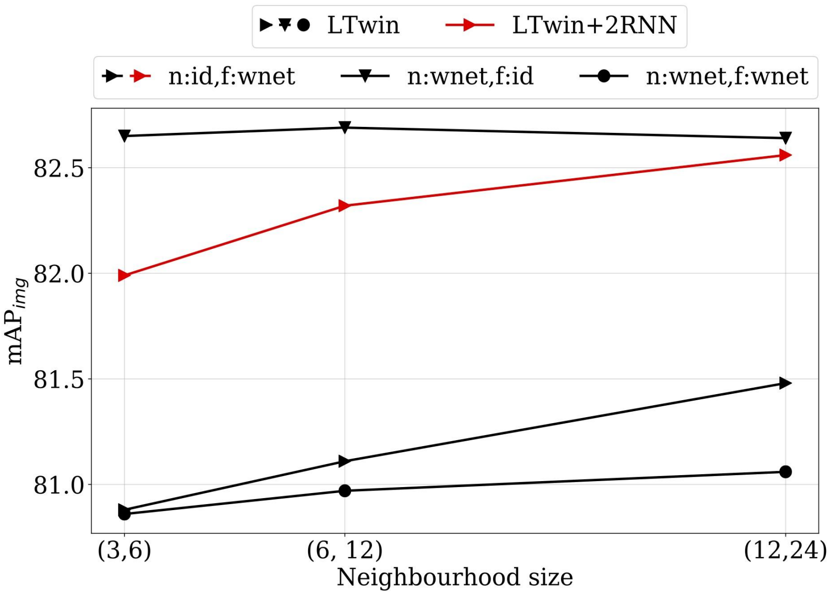




Experimental Results (4/4)

- mAP_{lab} and mAP_{img} for joint models (wordNet embeddings).





ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

quantexa
Connecting Data | Empowering Decisions

Thank You!

DAY 2 - January 13, 2021
Poster Session (PS) T3.6

Università degli Studi di Padova

tobiatesan@quantexa.com, {pasquale.coscia, lamberto.ballan}@unipd.it